Assisted curation of growth conditions that affect gene expression in E. coli K-12

Gama, Socorro -Castro ; Rinaldi, Fabio ; et al

Assisted curation of growth conditions that affect gene expression in *E. coli* K-12

Socorro Gama-Castro[1], Fabio Rinaldi[2], Alejandra López-Fuentes[1], Yalbi Itzel Balderas-Martínez[1], Simon Clematide[2], Tilia Renate Ellendorff[2], Julio Collado-Vides[1]

[1] Centro de Ciencias Genómicas, Cuernavaca, UNAM, México.

[2] Institute of Computational Linguistics, University of Zurich

Introduction

RegulonDB [4] is a database with manually curated knowledge extracted from the literature describing knowledge of transcriptional regulation in *E. coli* K-12. It contains objects such as genes, promoters, transcription factor binding sites (TFBSs), transcription factors (TFs), terminators and operons. It contains relations among those objects, such as regulatory interactions among TFs and genes, promoters and operons. An important piece of information for the adequate description of knowledge on gene regulation is that of growth conditions (GC) and their corresponding control conditions (CC), which are used in experiments to identify regulatory interactions. Currently RegulonDB has only a small set of GCs, which are known to activate or repress the transcription of a few genes. A list of the mechanisms by means of which the GCs affect gene expression is still missing.

The process of curation of the GC would require keeping track of a large amount of data about the experiment, such as the name of the GC, the control of the experiment, the growth media used, the temperature, the pH, the type of effect (induction or repression) provoked by the transition from CC to GC to the regulated gene, the TF and sigma factor involved, when known, in such regulatory mechanism. Thus, the biocuration challenge we face is to extract this type of relevant information from the large corpus of around 5,000 papers, supporting the knowledge of mechanisms present in RegulonDB with experiments performed since the 80s or even 70s to date. Doing this work manually would involve a considerable amount of time. This challenge motivated us to initiate our collaboration with experts in text mining tools, and use resources such as OntoGene/ODIN, to simplify and as such accelerate our curation.

The goal of this project is to verify which GCs activate or inhibit the transcription of the genes of *E. coli*, as well as to identify the type of mechanism used. In a first instance we can determine the type of mechanism based upon the identification of the TF and the effect it causes on some of the regulated genes under the given GC. Therefore we will try to identify the name

of the experimental condition, the affected gene, the type of effect, and the TF involved in such regulatory process.

OntoGene/ODIN provides a flexible, customizable environment for document-centric curation approaches. The OntoGene team at the University of Zurich, working in collaboration with RegulonDB curators, adapted ODIN to the specific needs of this project. Ontogene/ODIN has been previously described in several publications [1-3]. In the rest of this short paper we describe the results of the experiment on curation of GC for RegulonDB using ODIN.

Methods

We used the complete list of genes of *E. coli* from RegulonDB for building dictionaries to be used by OntoGene/ODIN. Additionally RegulonDB provides words that indicate the type of effect caused under a given GC (*activation*, *repression* and a complete list of their synonyms).

Our initial work has been performed on a set of 46 articles from RegulonDB that were selected because of their connection with the genes related to the regulon of OxyR, and with the regulatory interactions, operons, promoters and terminators of those genes. The articles have been automatically annotated by the OntoGene pipeline using the terminology provided by RegulonDB, which includes types such as GENE, EFFECT, Transcription Factors (TF), etc.

We use the sentence filters of ODIN to visualize, in those 46 articles, only those sentences containing the name of a GENE and a word of type EFFECT. Since we know that OxyR is a TF which is involved in the regulation of genes which respond to oxidative stress, we expect to find relevant data about GC in that set of articles.

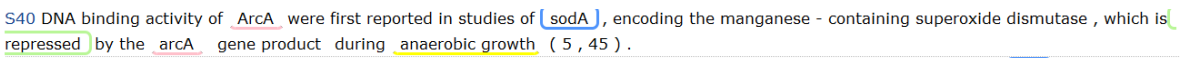
Since we have only an incomplete list of GC we cannot use the elements of the list as a filtering criteria to select relevant sentences in ODIN, since such a choice would severely limit the results. Our goal is in fact to discover the possible names and synonyms for GC. Because of that reason, after applying the filter, we use the automatically annotated genes and effects, but we manually mark previously missing GCs, in order to generate an extended set of such conditions.

Results

There were 36 out of the 46 articles with at least one sentence containing GCs-related information, which show the effect of a GC on the expression of at least one gene (see example in figure 1). Of these 36 articles, 20 contain at least one sentence that describes the mechanism of regulation at work under the specified GC. See an example in figure 2.

Figure 1

Figure 2



Other types of sentences found are those containing only information about the regulation of a gene by a TF (Regulatory Interaction), without mentioning the GC.

These results show that ODIN is a very useful instrument to help in the manual curation of RegulonDB. Some observations made during this experiment will help generate improved versions of the tools and terminological resources. For example, since all TFs are also genes, they received a duplicate annotation. However, if we want to curate only GC-related sentences, it would be better not to include the TFs in the list of genes used as a filter, because the terms related to an EFFECT, which are found in a sentence with GC, are also found in the sentences that contain only information of regulatory interactions. We also encountered GC-related data that regulate the activity of the TFs, and their mechanism, even if not necessarily at the level of transcription. This is also useful information for RegulonDB.

We spotted some errors in the automated annotation of some gene names. In particular short words (typically 4 letters or less) might also happen to be gene names. For example, “fold” is frequently used to express the level of expression of a gene, rather than to refer to the gene of the same name. Such errors were manually corrected.

A long-term goal is to use the system for a more specific, accurate and efficient curation. In the process of the experiment described above, we realized that it would be useful to be able to distinguish interrogative or hypothetical sentences from affirmative ones, since only the latter provide reliable data for curation. Another problem to solve is the lack of clarity when a mutant is mentioned in a sentence, without a description in the same sentence, since the sentence-based curation approach hides the information needed for complete understanding (being it contained in a non-selected sentence). A similar problem is caused by anaphoric mentions such as “this gene”, where the actual name of the gene is mentioned in a previous sentence, which might not be shown when the filter is active.

As soon as we have a complete list of GC, we will be able to use the ODIN sentence filters to allow a very detailed inspection of the documents and obtain more specific results.

An additional result of this practical experiment was to collect different ways in which GCs terms are described in articles. For example, stress conditions with hydrogen peroxide and exponential phase are written in different phrases with different words, such as:

- *H202, H2O2 exposure, H2O2 stress, H2O2 treatment, H2O2 - stressed cells, hydrogen peroxide, presence of hydrogen peroxide, exposure to hydrogen peroxide, hydrogen peroxide treatment, high concentrations of hydrogen peroxide, treated with hydrogen peroxide, treatment with hydrogen peroxide*

- *during growth, exponentially growing, exponential growth, exponential phase, exponentially growing cells, logarithmic – phase, log phase*

Analsys of the paper: PMID 21908668

The article with PMID 21908668 was analyzed in detail. The sentence splitter currently used in the OntoGene system identifies 841 sentences in this paper, although without the references there are only 327 of them. When the “GENE and EFFECT” filter was applied, 78 sentences were selected (three of them are part of the references).

From the total of 78 sentences: 18 sentences describe the regulation of a TF on a gene (TF-gene regulation) under a specific condition (TF-gene-GC); three sentences describe TF-gene regulation as well as the regulation of a GC on a gene (GC-gene regulation), although generally it is the same gene in both cases, it is not shown clearly the dependence of the GC with the TF. 13 sentences describe only GC-gene regulation. 12 sentences describe only TF-gene regulation. Five sentences are not clear, because they are questions, and not affirmative or negative sentences; 13 sentences contain general data about TF, e.g., it could be mentioned that a such TF is regulating a set of genes, but it is not specified to which genes; and finally 14 sentences do not contain the expected data, for example this group could have another kind of regulation where a TF or a GC is not involved.

In summary:

18 TF-gene- GC
3 TF-gene + GC-gene
13 GC-gene
12 TF-gene

5 Confusing sentences
13 TF
14 Nothing

The sentences that we expect to find belong to the first five groups, they represent 65% of the sentences when they are filtered. We also have the option to eliminate the TF group (with 13 sentences) if we exclude from the list of genes the names of the TFs. This would change the relevant set of sentences from 65 to 78%.

Conclusion and future work

The experiment clearly shows that efficient text mining tools coupled with a customized interface can significantly increase the efficiency and productivity of specific biocuration activities. The activity described in this paper required about 20 hours for the curation of 46 papers. Since the normal curation process at RegulonDB requires about four hours per paper, it can be estimated that the same activity, without the support of OntoGene/ODIN, would have required about 184 hours. It seems therefore that the careful introduction of sophisticated text mining and curation tools can improve the efficiency of curation nearly 10-fold.

We intend to continue the activities described in this paper within the scope of a planned collaborative project with curators of RegulonDB group and with the support of the OntoGene team. The goal is to gradually automatize much of the most tedious activities of the curation process, and therefore free up the creative resources of the curators for more challenging tasks, and enabling a much more efficient curation process.

References

1. Fabio Rinaldi, Thomas Kappeler, Kaarel Kaljurand, Gerold Schneider, Manfred Klenner, Simon Clematide, Michael Hess, Jean-Marc von Allmen, Pierre Parisot, Martin Romacker, Therese Vachon (2008). OntoGene in BioCreative II. Genome Biology, 2008, 9:S13, PMC2559984

2. Fabio Rinaldi, Simon Clematide, Yael Garten, Michelle Whirl-Carrillo, Li Gong, Joan M. Hebert, Katrin Sangkuhl, Caroline F. Thorn, Teri E. Klein, Russ B. Altman. *Using ODIN for a PharmGKB revalidation experiment. The Journal of Biological Databases and Curation*, Oxford Journals, 2012, bas021; doi:10.1093/database/bas021

3. Fabio Rinaldi and Simon Clematide and Simon Hafner and Gerold Schneider and Gintare Grigonyte and Martin Romacker and Therese Vachon. Using the OntoGene pipeline for the triage task of BioCreative 2012. The Journal of Biological Databases and Curation, Oxford Journals, 2013. doi:10.1093/database/bas053

4. Salgado H, Peralta-Gil M, Gama-Castro S, Santos-Zavaleta A, Muñiz-Rascado L, García-Sotelo JS, Weiss V, Solano-Lira H, Martínez-Flores I, Medina-Rivera A, Salgado-Osorio G, Alquicira-Hernández S, Alquicira-Hernández K, López-Fuentes A, Porrón-Sotelo L, Huerta AM, Bonavides-Martínez C, Balderas-Martínez YI, Pannier L, Olvera M, Labastida A, Jiménez-Jacinto V, Vega-Alvarado L, Del Moral-Chávez V, Hernández-Alvarez A, Morett E, Collado-Vides J. RegulonDB v8.0: omics data sets, evolutionary conservation, regulatory phrases, cross-validated gold standards and more. Nucleic Acids Res. 2013 Jan;41(Database issue):D203-13. doi: 10.1093/nar/gks1201. Epub 2012 Nov 29.